Developments in Standardised Assessment:

A Perspective from the UK

Christine Merrell,
Centre for Evaluation and Monitoring (CEM),
Durham University

I begin this chapter by considering the purposes of assessment in childhood. I then focus on reasons for standardised assessment. Within the definition of standardised assessment, I include the way that assessments are conducted and aspects of the standardisation of outcomes in terms of norm referencing. Examples of standardised assessments and their use are considered, ranging from the more traditional approach of pencil and paper group assessments to recent developments in personalised computer-delivered assessments. I then discuss the merits and disadvantages of different methods and reach the conclusion that it is not only the type of assessment that matters but how it is used.

PURPOSES OF ASSESSMENT IN CHILDHOOD

We are assessed before we are even born. Expectant mothers in many countries are given ultrasound scans at various stages of pregnancy to check that the foetus is developing normally and to estimate when the baby is likely to be born. Immediately after birth, a baby is weighed and his or her condition and reactions are scored against standardised scales to decide whether specialist care is required. The assessments continue as the baby grows with, for example, the Newborn Hearing Screening Programme which measures the hearing of 99% of all babies born in England during the first few weeks of their lives so that early intervention can be offered where needed. Assessments continue through childhood and into adulthood.

In terms of educational assessment, there are many different reasons for collecting information. Newton (2007) presented a list of eighteen purposes for which educational judgements may be used. He acknowledged that the list was not exhaustive and that the purposes for which information derived from assessments is used are expanding all the time.

If individuals or groups are going to be compared to each other, we need to be sure that they have all been assessed in the same way; namely, the content of the assessment, the way that is scored and the interpretation of those scores should be the same for all individuals. In other words, the assessment is standardised.

Different users of assessment data require different levels of detail. From the beginning of kindergarten, a child's teacher needs to have high-quality information about what children can do and understand, including their cognitive development, personal, social and emotional development, motor development, dispositions and

attitudes, in order to plan and support appropriate learning experiences. At this early stage, potential special educational needs may be identified. This information is useful to be able to appropriately target resources, although a formal diagnosis of a specific difficulty should be made with caution, bearing in mind that children develop at different rates and difficulties may simply be due to immaturity. An example of problems which may be a consequence of immaturity rather than a chronic disorder is the behaviour of young children. Many young children have a limited span of attention, are very active and impulsive. These behaviours, however also characterise the diagnostic criteria for Attention Deficit Hyperactivity Disorder (ADHD), thought to be a consequence of impaired behavioural inhibition and executive functions (Barkley, 1997). Importantly, very few children with such 'symptoms' are eventually diagnosed with ADHD because behavioural inhibition and executive functions continue to develop throughout childhood. Therefore even accurate assessment does not necessarily reflect accurate diagnosis.

Assessments conducted in a variety of ways and from different sources, including a child's teacher and parents, can build up a comprehensive profile which can be shared between them to provide complimentary care, education and support for personal development. At an individual level, assessments administered at regular intervals throughout kindergarten and school can monitor the amount of progress that children are making in their development, knowledge and skills.

Aggregated assessment scores provide information about groups of children from which comparisons can be made. Aggregated information can be useful to a number of stakeholders; a teacher who wishes to obtain a broad picture of the strengths and weaknesses of the class as a whole, school managers who might wish to compare classes and cohorts, district officers who are making comparisons between schools, or by policy-makers at national level. These comparisons, from an individual to population level, are all valid and important uses of assessments and require information that is collected using reliable standardised assessments of individual children at regular intervals.

Comparisons of standards within a single cohort may be made from assessments administered to groups at a particular time in their education. For example, an assessment administered to all pupils in a country at the end of a stage of learning or upon completion of compulsory education can be used to compare the performance of students in different schools or districts across a country. Large-scale, international studies compare educational standards across the world.

Assessments which remain comparable in their content from one year to the next can be used to monitor changes in standards over time. For this type of comparison, not only is it necessary to have an assessment whose *content* is comparable over time, it is also necessary to conduct the assessment with *groups* that are comparable over time, such as students who took an assessment at the same age or time of year. For example, Merrell and Tymms (2010) investigated changes in children's cognitive development on entry to school in England over a period of nine years from 2001 to 2008. We collected data from every child starting school, each year, in the same sample of four hundred and seventy schools. The children were assessed by their class teachers within the first six weeks of the academic year using a computer-delivered assessment of early reading, vocabulary and mathematics, which remained the same

over the period. Studies such as these reflect the impact of government policies and contribute a perspective to inform future direction. The period covered by the Merrell and Tymms' study was a time of significant investment (tens of billions of GBP) by the English government into pre-school facilities to improve the educational outcomes of children in the early years, particularly those children from deprived backgrounds. If the interventions had been effective, it would have been expected that children's early vocabulary, reading and mathematics development at the start of school would have improved. The study showed that there was no difference in children's vocabulary and reading standards and very little improvement in mathematics. Interestingly, although the Government commissioned a national evaluation of those policies, it had not implemented the necessary standardised assessments to assess their impact on a national scale.

Standardised *assessments* enable comparisons between individuals and groups to be made. If they are administered to a representative sample or full population, standardised *scores* can be calculated.

So far, some of the purposes of assessment and uses of information from standardised assessments have been discussed in general terms and examples given. The next section of the chapter explores different stages of schooling and why it is useful to measure their impact in a standardised way.

BASELINE ASESSMENT ON ENTRY TO SCHOOL

There is a long tradition of conducting a baseline assessment of children's development on entry to school. Within England, back in the 1960s, the main purpose of a baseline assessment at the start of school, reflected by the popular instruments of the time, was to identify children's special educational needs at an early stage rather than establishing a baseline from which progress in schooling could be measured.

In 1998, there was a change in focus as on-entry baseline assessment of children within the first few weeks of starting school became a statutory requirement for all English schools which received any state funding (Blatchford & Cline, 1992; Wolfendale & Lindsay, 1999). Establishing a reliable baseline at the start of formal education from which progress can be monitored has the potential to identify individuals, classes, schools and districts where progress is lower than expected and thus remedial actions can be taken. Over ninety baseline assessment schemes were accredited and schools were able to select the scheme of their preference.

One such accredited scheme, which had the biggest market share, was the Performance Indicators in Primary Schools On-entry Baseline Assessment (PIPS-BLA), was published by the Centre for Evaluation and Monitoring (CEM) at Durham University, UK (Tymms, 1999). The content of the PIPS BLA was underpinned by research (Tymms, 1999) and it was used on a large scale with mainstream children to provide reliable information for teaching and learning. It includes measures of vocabulary acquisition, concepts about print, phonological awareness, letter and word recognition, reading and comprehension, concepts about maths, digit identification and simple number problems. These were all areas of development which were identified from published research as being strongly related to later educational outcomes. The PIPS BLA is a computer-adaptive assessment which is administered

by an adult working with one child at a time. The software tailors the assessment to the ability of a child on the basis of his or her answers and the whole assessment takes between fifteen and twenty minutes per child. The method of administration, which combines the presentation of items by the computer with teachers' decisions on whether children answer correctly or incorrectly, is standardised to such a degree that the re-test reliability has been found to be 0.98 and the internal reliability, measured by Cronbach's alpha, is 0.94 (Tymms, Merrell & Jones, 2004; Tymms, 1999). The assessment was originally developed for use by schools in the United Kingdom from which progress in the elementary years could be measured. In recent years, it has been adapted and, where necessary, translated for use in many other countries including Abu Dhabi, Australia, Germany, Hong Kong, the Netherlands, New Zealand and South Africa. After analysing the data for cultural bias, international comparisons of children starting school in different countries have been published (Tymms, Merrell & Jones, 2004). The large-scale international studies of student attainment (e.g. TIMSS) compare the effect of education across different countries but that information is limited without knowing the knowledge and skills which children started school with and thus the progress that they have made between that point and the later assessment. The study by Tymms et al. demonstrates that it is realistic to compare children's development at the start of school in different countries. Such a comparison reflects each country's policy on pre-school care and education as well as offering a context for the interpretation of the data collected in later years by the international studies. The authors compared children starting school in Australia, England, the Netherlands, New Zealand and Scotland. Although the study included several thousand children, not all countries' samples were nationally representative therefore it was proposed as a pilot study which demonstrated a model for a much larger, systematic study. A linear relationship between age and reading/maths development was found at the start of school and countries largely fitted on that line. Deviations were found for sub-groups for example the scores of the indigenous children in Australia were consistently lower for their age compared with other children.

A reliable assessment of children's development at the start of school is crucial for teachers to plan appropriate learning experiences, but standardised scores which compare children's development against population norms are, for that purpose, less important. However, when it comes to identifying learning difficulties or gifted children for, perhaps, the allocation of scarce resources or specialist help, standardised assessments provide an extremely useful reference point. Although some teachers will have extensive experience working with a wide range of children in different situations and can spot deviations from the norm that warrant specialist intervention, many do not. Without the reference to population norms, how would those teachers with less experience of a wide range of children know whether or not a child's development was significantly different to the average?

MONITORING PROGRESS IN SCHOOL

Moving on from an initial assessment of children at the start of school, teachers need feedback about how pupils are progressing. In her book titled 'Monitoring Education: Indicators, Monitoring and Effectiveness' (1997), Carol Taylor Fitz-Gibbon described the work of a number of statisticians and researchers, for example W. Edwards Deming, who followed the principal of identifying a problem, proposing and implementing a solution, monitoring the impact of that solution and adjusting as

necessary. Although these methods had been applied to processes such as engineering and production, and while acknowledged the complexity of education, she nevertheless suggested that they were applicable to educating children.

Of course the processes can be applied to the growth and development of individual children using assessments which are not standardised, and in this way teachers would be able to see the value that education has added. However, although teachers can see their pupils learning new things and developing, without standardised assessments at regular points throughout their education it is difficult to estimate whether a child is making good progress compared with others of the same age, ability and time in school, that is, whether their methods are as effective as those of other teachers. This method of feedback, where a baseline measure is used to predict an outcome measure for a group, and then individuals' performance is measured against the group's regression line can be described as 'relative value-added'. That is, a comparison of how well individual children, classes, schools or districts are progressing in comparison with others. It is from employing these statistical methods to the analysis of standardised assessments that the area of school effectiveness has grown (for an overview and history of school effectiveness research, see Teddlie & Reynolds, 2000).

Carol Taylor Fitz-Gibbon initially set up large-scale monitoring systems for pre-university courses and these were extended by her and Peter Tymms to provide primary and secondary schools with measures of relative value-added feedback about their pupils (Taylor Fitz-Gibbon (1996); Tymms (1999); Tymms & Albone (2002). These systems are run by the Centre for Evaluation and Monitoring (CEM) at Durham University, England. CEM's monitoring systems assess children at multiple points in their schooling, thus providing trajectories of growth. They were originally used by schools in England but have expanded into several countries (www.cemcentre.org). Taylor Fitz-Gibbon advised the English government on setting up a national value-added system for its schools (Taylor Fitz-Gibbon, 1997). Similar models have evolved across the world, many in the USA (see, for example, Sanders and Horn, 1994). Value-added systems can pose a threat to teachers and head-teachers if the results are made publicly available in the form of published league tables of the type seen in the English media in recent years. This inevitably leads to stress within the profession and indeed of pupils themselves, a narrowing of the curriculum with teachers focussing entirely on the content of the tests and the de-motivation of pupils (National Union of Teachers, 2006). Teachers and schools do need high quality information but there needs to be an element of public trust and respect that this is being used in a professional way to improve pupils' outcomes.

Value-added feedback is also important for evaluating the impact of interventions and policies at a national level. For example, does a national policy to teach children to read using a systematic synthetic phonics programme significantly raise reading levels? Without a standardised assessment system to analyse the progress made by children, it is difficult to know. Once the impact of an intervention on pupils' outcomes has been established in terms of effect size (see for example Coe, 2002), this can be compared against the impact of other interventions and cost-benefit calculations performed.

Another use for standardised assessments is to investigate the importance of teachers, head teachers and districts with respect to pupils' progress. Effectiveness studies have tended to focus on the school as the unit of analysis but a study by Tymms et al. (2008) used standardised data to compare the effectiveness of districts, schools and teachers. The authors found that the district in which a child attended a school made virtually no difference to the amount of educational progress, the school made more of a difference but the most influential factor was the teacher. Knowing that it is effective teachers (see for example Nye, Konstantopoulos and Hedges, 2004) which make a difference to children's progress and outcomes rather than the head teacher or input from the district can influence decisions about resourcing in schools.

ADVANTAGES AND LIMITATIONS OF STANDARDISED ASSESSMENTS

Having discussed some of the uses and values of standardised assessments, I now explore issues associated with different methods of conducting them.

Assessments of children can be made using a variety of methods. For example, if we wish to find out if children can perform a mathematical calculation or if they can read high-frequency words, we can ask them questions verbally, by computer or using the traditional pencil-and-paper format. Questions which have a single, defined, correct answer are relatively easy to mark with virtually no judgement about the quality of the response required. However, it is not appropriate to assess all areas of development by direct questioning. If we wish to learn more about a young child's behaviour, for example how well they interact with their peers or whether they can demonstrate sustained attention during a task, observing them in a natural setting over a period of time is likely to give a more reliable and valid result.

Each method has advantages and problems, yielding different amounts of information to different degrees of reliability, comparability and validity. Short questions with multiple choice answers are quick to administer and marking can be automated. Typically, tests of this format include many items and have high internal reliability but they have limitations; children are more likely to get an answer correct by chance than with constructed response items, they are often found to be biased towards boys (see for example Ben, Shakhar, & Sinai, 1991) and since the answer options are limited in the information that they present, their focus can be narrow and not always elicit children's full understanding of concepts.

Assessment items requiring a constructed response, which is judged against a set of criteria, also have limitations. The judgements must be standardised if the results are going to be comparable. Newton (2009) investigated the reliability of the results from statutory national curriculum tests completed in England. He found that the marking of the mathematics test was the most reliable followed by science and, considerably lower was writing. The mathematics test answers required the least interpretation by markers and the writing the most.

Assessing children's knowledge through observation alone has reliability issues. Firstly, a child may have sophisticated, in-depth knowledge but does not display it without being prompted. A young child might know how to perform complex mathematical computations but may be too timid to demonstrate that in classroom

activities without prompting, or they may simply choose not to. An example of an assessment of children's knowledge and understanding that, it is suggested, should be conducted predominantly through observation of child initiated activities is the Early Years Foundation Stage Profile (EYFSP), which is a statutory assessment of children in Early Years Foundation Stage settings in England covering six areas of learning. The guidance for practitioners on completing the Profile stated that:

> "Observational assessment is the most effective way of making judgements about all children's development and learning." (QCA, 2008, Page 14)

> "Practitioners need to ensure that they are observing children as a key way of understanding what they really know and can do. This is demonstrated most effectively when children are engaged in self-initiated activities. Because self-initiated activities will take place within provision in which the adults have made decisions about which resources and equipment are available, it is important to clarify the definition of this." (QCA, 2008, Page 9.)

Yet the guidance document does not provide evidence from well conducted experiments to support the view that observations of self-initiated activities give more reliable information about what a child really knows and can do than other methods of assessment. At the present time, the assessment includes 117 scale points for which observations must be made and I would argue that to assess a class of up to twenty five children in the recommended method is not an effective use of teachers' time and it not does it necessarily provide teachers with detailed information about their pupils' strengths and areas for development. It takes time to make detailed observations about a class of children and during that time the learning experiences that are afforded them might not be sufficiently tailored to their zone of proximal development. A recent independent review (Tickell, 2011) has called for the number of scale points in the EYFSP to be significantly reduced.

There is another problem which the guidance acknowledges:

> "There are some groups of children for whom this challenge needs particular consideration so that their attainment is not underestimated." (Page 14)

The groups of children which are listed are: (a) those with English as an additional language, (b) boys, (c) children with special educational needs and (d) children from minority groups. This is more than half of the population and a contradiction to the earlier statement that observational assessment is the most effective way of making judgements about all children's development and learning.

Problems with teachers' judgements are widely documented. Harlen (2005) suggested caution with regards to teachers' judgements of pupils' attainment and progress:

> "The findings of the review by no means constitute a ringing endorsement of teachers' assessment; there was evidence of low reliability and bias in teachers' judgements."

Harlen referred to bias and this can be introduced when an assessor systematically downgrades an individual or group for construct irrelevant reasons. Several instances of bias in teacher assessments have been documented in relation to several factors, for example sex, ability, ethnicity, social class, age and behaviour. Harlen's systematic

literature review of the evidence of reliability and validity of assessment by teachers used for summative purposes (2004) and Wilmut's investigation of the experiences of summative teacher assessment in the UK (2005) synthesised the findings of many studies.

A further example of bias in teachers' ratings in relation to children's ethnicity comes from a study by Sonuga-Barke, Minocha, Taylor and Sandberg (1993). The authors investigated the relationship between teachers' ratings of hyperactivity and attention in groups of children classified as being of Asian or English origin, attending primary schools in one London Borough. Teachers completed questionnaires and structured interviews to rate their pupils' behaviour. At the same time, objective measures of the activity and attention were taken. The teachers judged their pupils of Asian origin to be more inattentive and hyperactive than their peers of English origin. However, there was a discrepancy between the teachers' judgements and the scores derived from the objective measures for the children of Asian origin. The objective measures suggested no significant difference between the groups. The authors concluded that teachers appeared to over estimate the Asian children's levels of activity relative to those of the English children; a bias in their ratings.

The examples described so far have considered bias with groups of children in relation to those group characteristics which can be irrelevant to the construct being measured. Another form of bias, sometimes referred to as the Halo Effect, occurs within subjects and can be a feature of the outcomes of assessments which are conducted through observation alone. Examples have been found across disciplines and are not limited to educational assessments (see Scorcher and Brant (2002) and Rosenzweig (2007) for examples in business and leadership). The Halo Effect occurs in educational assessment when a teacher rates a child as being competent in one subject area and because of the impression formed, she will also tend to rate the child as being equally competent in other areas such as motor development or personal, social and emotional development.

One example where the Halo Effect is evident is within the EYFSP described earlier. To illustrate this, one hundred and six pupils' EYFSP scores from the six areas of learning which it covers (Language, Mathematics, Personal, Social and Emotional Development (PSED), Knowledge and Understanding of the World (KUW), Motor Development and Creative Development (Create Dev.) were compared against each other and against their scores from the PIPS Baseline Assessment (described earlier) for Language and Mathematics. These children were assessed in the 2002/03 academic year. The EYFSP was conducted solely through observation and PIPS was conducted by asking each pupil questions in a standardised way by a computer-delivered program. The scores from the end of the first year at school of one hundred and six children in three English primary schools were analysed and the correlations were as follows:

	EYFSP PSED	EYFSP Lang.	EYFSP Maths.	EYFSP KUW	EYFSP Motor Dev.	EYFSP Create Dev.	PIPS Lang.
EYFSP Lang.	0.81						
EYFSP Maths.	0.84	0.92					

EYFSP KUW	0.84	0.83	0.83				
EYFSP Motor Dev.	0.83	0.72	0.77	0.75			
EYFSP Create Dev.	0.80	0.82	0.79	0.85	0.81		
PIPS Lang.	0.65	0.82	0.80	0.74	0.61	0.66	
PIPS Maths	0.66	0.72	0.76	0.65	0.62	0.64	0.82

All correlations were significant at the 0.01 level (2-tailed).

The correlations between EYFSP sections are all high. These areas of development are correlated to a certain extent, with the strongest association generally found between language and mathematics. The association between language or mathematics and motor development is generally weaker, for example Son and Meisels (2006) found correlations of 0.4 between the reading and visual motor skills, and 0.2 between the reading and gross motor skills of children in kindergarten when children were assessed with objective measures. Yet whilst the correlations between the EYPSP language and mathematics are high, as expected, the correlations between the EYFSP motor development and language or maths scores are much higher than expected. The correlations between the PIPS measures and EYFSP are high for language and maths, as would be expected, but lower between PIPS and the other EYFSP measured. This demonstrates a Halo Effect occurring within the EYFSP.

The EYFSP is an example of an assessment method which is at one extreme, relying predominantly on the observation of child-initiated activities, but the assessment of children's attainment and progress using a combination of teacher assessment and standardised objective assessments for the purpose of accountability is becoming a more widespread feature of national systems. Scotland has recently changed its national assessment system to align with its new curriculum (Curriculum for Excellence). In the guidance on assessment issued to all Scottish schools and councils (The Scottish Government, 2010), a combination of teacher assessment by comparing pupils' attainment against national exemplars of standards along with standardised objective assessments is recommended. More recently, in 2011, the English government commissioned an independent review of testing arrangements at the end of the primary phase of education (referred to as the end of Key Stage Two). Evidence from many sources, including expert opinion, was gathered and the recommendation made to include a greater element of teacher assessment than currently occurs. A further recommendation was to consider the use of computer-delivered assessments, including diagnostic computer-adaptive programs (Bew, 2011).

Diagnostic computer-adaptive assessments are becoming more widespread and hold promise by producing detailed information about children's strengths and areas of difficulty that can be used to develop personalised education[1]. They are motivating for children, presenting questions that are within their zone of proximal development, and produce a more reliable measure especially for children at the high or low end of the

[1] For a description of the development of a computer-adaptive diagnostic assessment of reading, see Merrell and Tymms (2006), and for a wide-ranging discussion of computer-adaptive testing for reading, see Chalhaub-Deville (2000)

ability range. I have given examples of the computer-adaptive assessments produced by CEM at Durham University, which are used on a large scale with hundreds of thousands of children using them each year and questions could be asked about them discriminating against children with additional support needs or with limited experience of ICT. We have investigated these possible issues and not found evidence of their presence. Moreover, we have received positive feedback about particular groups of children for example the indigenous children in Western Australia who, at the time of the report, had not been exposed to ICT-rich environments. Teachers reported the children's delight at the assessment format. Of course, as with any standardised assessment, care must be taken to eliminate bias and discrimination including ensuring fair access for children with a range of additional support needs such as sensory impairments or attentional difficulties. These assessments are valuable and flexible standardised tools whose uses are only just beginning to be realised in terms of their power to provide instant norm and criterion referenced feedback to both pupils and teachers.

CONCLUSIONS

Standardised educational assessment has many benefits and the chapter has discussed some examples. There are issues with how standardised assessments are conducted and despite the concerns expressed about observation methods, teachers do need a full picture of a child if they are going to cater for all their needs and this picture cannot realistically be achieved through objective assessments alone. A balance of assessment methods is required and there will always be a particular aspect of a child's behaviour or development elicited through a non-standardised method which is useful and important within the context of that individual. It is not only the type of assessment that matters but the way it is used.

Looking to the future, recent advances in computer-adaptive diagnostic assessments have the potential to provide efficient and reliable group assessments which probe children's knowledge and understanding in a detailed and appropriate way that has not been possible with more traditional group assessment methods, and to provide rapid feedback for improvement that is tailored to an individual's needs.

References

Barkley, R. A. (1997). Behavioural Inhibition, Sustained Attention, and Executive Functions: Constructing a Unifying Theory of ADHD. *Psychological Bulletin, 121*, 65-94.

Ben - Shakhar, G. and Sinai, Y. (1991) Gender Differences in Multiple Choice Tests: The Role of Differential Guessing Tendencies, *Journal of Educational Measurement*, *28*, 23-35.

Bew, Lord (2011) *Independent Review of Key Stage 2 testing, assessment and accountability: Final Report*. www.education.gov.uk Accessed 02 July 2011.

Blatchford, P. & Cline, T. (1992). Baseline assessment for school entrants, *Research Papers in Education*, *7*, 247-269.

Chalhoub-Deville, M. (Ed.) (2000). *Issues in Computer-Adaptive Testing of Reading Proficiency: Studies in Language Testing 10*. Cambridge: Cambridge University Press.

Coe, R. (September, 2002) It's the Effect Size, Stupid: What Effect Size is and why it is important, Paper presented at the Annual Conference of the British Educational Research Association, Exeter, England.

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In the Assessment and Learning Research Synthesis Group *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2005). Trusting teachers' judgement: research evidence of reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*, 245-270.

Merrell, C., & Tymms, P. (2006) Identifying Reading Problems with Computer-Adaptive Assessments. *Journal of Computer Assisted Learning*, *23*, pp.27-35.

Merrell, C., & Tymms, P. (2010) Changes in Children's Cognitive Development at the Start of School in England 2001 – 2008. *Oxford Review of Education*. iFIRST 1-13, http://www.informaworld.com/smpp/title~content=t713440173

National Union of Teachers (2006) Briefing: The Impact of National Curriculum Tests on Pupils.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*, 149-170.

Newton, P. E. (2009). The reliability of results from national curriculum testing in England, *Educational Research*, 51:2, 181-212

Nye, B., Konstantopoulos, S., & Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26,* 237–257.

Qualifications and Curriculum Authority (2008) *Early Years Foundation Stage Profile Handbook*, Pub. Her Majesty's Stationery Office: London

Rosenzweig, P. (2007). The Halo Effect … and the Eight Other Business Delusions that Deceive Managers. *Free Press*.

Sanders, W. & Horn, S. (1994) The Tenessee Value-added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment. *Journal of Personnel Evaluation in Education, 8,* 299-311.

Scorcher, M., & Brant, J. (2002). Are You Picking the Right Leaders? *Harvard Business Review*, February.

The Scottish Government, (2010). *Curriculum for Excellence; Building the Curriculum 5; A Framework for Assessment.* Edinburgh: The Scottish Government.

Sonuga-Barke, E. J. S., Minocha, K., Taylor, E.A., & Sandberg, S. (1993). Inter-ethnic bias in teachers' ratings of childhood hyperactivity. *Journal of Developmental Psychology, 11,* 187-200.

Son, S-H, and Meisels, S.J. (2006), The Relationship of Young Children's Motor Skills to Later Reading and Math Achievement, *Merrill-Palmer Quarterly,* 52(4), 755 – 778.

Taylor Fitz-Gibbon, C. (1996). *Monitoring Education: Indicators, Quality and Effectiveness*. Continuum: London.

Taylor Fitz-Gibbon, C. (1997). *The Value Added National Project: Final Report: Feasibility Studies for a National System of Value Added Indicators*. London: SCAA.

Teddlie, C., & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research.* Falmer Press: London.

Tickell, C. (2011). The Early Years Foundation Stage (EYFS) Review: Report on the Evidence.
http://media.education.gov.uk/assets/Files/pdf/T/The%20Tickell%20Review.pdf
Accessed 02 July 2011.

Tymms, P. (1999). *Baseline assessment and monitoring in primary schools: achievements, attitudes and value-added indicators*. London, David Fulton.

Tymms, P., & Albone, S. (2002). Performance Indicators in Primary Schools. *School Improvement Through Performance Feedback*. In A. J. Visscher & R. Coe. Lisse/Abingdon/Exton PA/Tokyo, Swetz & Zeitlinger (pp191-218).

Tymms, P., Merrell, C., Heron, T., Jones, P., Albone, S. & Henderson, B. (2008). The Importance of Districts, *School Effectiveness and School Improvement, 19,* 261-274.

Tymms, P., Merrell, C. & Jones, P. (2004). Using baseline assessment data to make international comparisons, *British Educational Research Journal*, *30*, 673 – 689.

Wilmut, J. (2005). *Experiences of summative teacher assessment in the UK; A review conducted for the Qualifications and Curriculum Authority*. London: QCA.

Wolfendale, S. & Lindsay, G. (1999). Issues in baseline assessment. *Journal of Research in Reading 22*, 1-13.